A large deviations principle for infinite-server queues in a random environment

H. M. Jansen[1,2], M. R. H. Mandjes[1], K. De Turck[2], S. Wittevrongel[2]

February 4, 2015

Abstract

This paper studies an infinite-server queue in a random environment, meaning that the arrival rate, the service requirements and the server work rate are modulated by a general càdlàg stochastic background process. To prove a large deviations principle, the concept of attainable parameters is introduced. Scaling both the arrival rates and the background process, a large deviations principle for the number of jobs in the system is derived using attainable parameters. Finally, some known results about Markov-modulated infinite-server queues are generalized and new results for several background processes and scalings are established in examples.

Keywords. infinite-server queue ⋆ random environment ⋆ modulation ⋆ large deviations principle

[1] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands.

[2] TELIN, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium.

E-mail. {h.m.jansen|m.r.h.mandjes}@uva.nl, {kdeturck|sw}@telin.ugent.be

1 Introduction

The infinite-server queue is one of the fundamental models in queueing theory. Its distinguishing feature is the presence of an infinite number of servers, so that jobs are served independently and there are no waiting times. This leads to explicit formulas for many quantities of interest, especially for M/M/∞ queues, where jobs arrive according to a Poisson process and the service requirements have an exponential distribution. In practice, however, one often observes time-varying arrival intensities, service requirement distributions and server work rates. This calls for adequate modeling.

A natural way to incorporate time-dependence is to consider an M/M/∞ queue in a random environment. In this case there is an independent background process that modulates the arrival rate, the service requirement distribution and the work rate of the servers.

Model. In this paper, we study the case where the background process is a general stochastic process J whose paths are right-continuous and have finite left limits, i.e. J has càdlàg paths. The process J modulates the arrival rate, the service requirement distribution and the server work rate in the following way. When J is in state x, jobs arrive according to a Poisson process with intensity $\lambda(x)$. Upon arrival, a job draws an independent service requirement from an exponential distribution with parameter $\kappa(x)$ if J is in state x when the job arrives. Then the service requirement of the job is processed by a server, whose work rate is $\mu(x)$ while J is in state x. Immediately after its service requirement has been processed, a job leaves the system.

Main result. The main result of this paper is a full large deviations principle (LDP) for the transient number of jobs in the system, under a scaling of the arrival rate and the background process. To arrive at this result, we first show that the number of jobs in the system at time $t \geq 0$ has a Poisson distribution with random parameter $\phi_t(J)$. Then we scale $\lambda \mapsto n\lambda$ and we scale $J \mapsto J_n$ such that the normalized random parameter $\phi_t(J_n)$ satisfies an LDP. Under this scaling, we derive the LDP for the transient number of jobs in the system.

Literature. The amount of literature on infinite-server queues in a random environment is quite small. Moreover, almost all papers on this topic (with notable exception [5]) study Model I or Model II (cf. [3]). In both models, jobs arrive according to a Poisson process with intensity $\lambda(x)$ when the background process is in state x. In Model I, service requirements have a standard exponential distribution and servers work at rate $\mu(x)$ when the background process is in state x. This is equivalent to the jobs being subject to a modulated hazard rate. In Model II, service requirements have an exponential distribution with parameter $\kappa(x)$ when the background process is in state x and servers work at constant rate 1.

An early reference is [14], which analyzes Model I when the background process is a continuous-time Markov chain. Important results in [14] are a recursion for the factorial moments of the number of jobs and the observation that the steady-state distribution is not of some 'matrix-Poisson' type.

Other important results can be found in [7], which studies Model I when the background process is a semi-Markov process with finite state space. The crucial observation in [7] is that the stationary number of jobs has a Poisson distribution with a random parameter that is determined by the background process. Moreover, the factorial moments of the number of jobs are computed via a recursion. These results are generalized in [12].

The observation in [7] is used to obtain time-scaling results in both the central limit regime and the large deviations regime. In the central limit regime, [2] and [4] derive central limit theorems for Markov-modulated infinite-server queues for several models and scalings. In this regime, the so-called deviation matrix (cf. [6]) plays an important role. In the large deviations regime, [3] and [5] compute optimal paths to obtain rate functions under a linear scaling of the arrival rates, given that the background process is an irreducible continuous-time

Markov chain. The former studies Model I, whereas the latter studies Model II for a class of service requirement distributions that includes the exponential distribution.

As mentioned, we show that the number of jobs in the system has a Poisson distribution with a random parameter, which can be interpreted as a mixture of Poisson distributions. In [1], an LDP is derived for mixtures that satisfy certain assumptions. However, apart from the assumption that the normalized random parameter satisfies an LDP, these assumptions are either superfluous or too restrictive in our case. In particular, we do not assume that the sequence of measures induced by the normalized random parameter is exponentially tight, so we cannot use the arguments in [1]. Hence, we need a different approach to obtain an LDP.

Contributions. In more detail, the contributions of this paper are the following. We generalize known models by considering a general càdlàg background process instead of a semi-Markov background process with finite state space. Moreover, in our model the background process modulates both the service requirement distributions and the server work rate, whereas previous papers considered models in which either the service requirement distributions or the server work rate was modulated. In particular, our model generalizes Model I and Model II.

Using elementary arguments, we show that in this model the transient number of jobs has a Poisson distribution with random parameter. We scale the arrival rate linearly and we scale the background process such that the normalized random parameter satisfies an LDP. Under this scaling, we obtain a full LDP for the number of jobs in the system. To the best of our knowledge, this is the first time that a full LDP is presented for modulated infinite-server queues. To prove the LDP, we introduce the concept of attainable parameters and use a variation on Varadhan's Lemma. These tools enable us to avoid the assumptions in [1].

The theory is illustrated by examples that show rate functions that cannot be obtained via background processes with finite state space. Additionally, we show that completely different background processes may lead to the same LDP, even in highly nontrivial cases.

Organization. The rest of this paper is organized as follows. In Section 2, we describe the model and provide some of its basic properties. Additionally, we fix some notation. In Section 3, we introduce the concept of attainable parameters and prove an LDP for the number of jobs in the system. In Section 4, we show that the rate function corresponding to this LDP has a simple description when we do not scale the background process. As an illustration, we work out some examples. In Section 5, we give examples in which we do scale the background process. In Section 6, we briefly discuss the results and point out some topics for future research. The appendices provide some technical details about the number of jobs in the system (Section A), continuity in Skorokhod space (Section B) and properties of Poisson random variables (Section C).

2 Model and problem description

We study an infinite-server queue with modulated arrival rates, service requirements and server work rates. The precise mathematical setup of the model and some of its basic properties are provided in Section A. In words, the model may be described as follows.

Let $(J(t))_{t\geq 0}$ be a càdlàg stochastic process with state space $\mathcal{E}$, which is assumed to be a metric space. We will refer to the process J as the background process or modulating process. While the background process is in state $x \in \mathcal{E}$, jobs enter the system following a Poisson process with intensity $\lambda(x) \geq 0$.

When job k enters the system, it draws a service requirement from an independent exponential distribution with parameter $\kappa(y)$ if the background process is in state $y \in \mathcal{E}$ upon its arrival. Server k processes this service requirement at rate $\mu(z)$ while the background process is in state $z \in \mathcal{E}$. Job k leaves the system when its service requirement has been processed.

We denote a modulated infinite-server queue by the quadruple $(J, \lambda, \kappa, \mu)$. Additionally, we denote the number of jobs in this system at time t by $M(t)$. In Section A it is shown that $M(t)$ has a Poisson distribution with random parameter

$$\phi_t(J) = \int_0^t \lambda(J(s)) e^{-\kappa(J(s)) \int_s^t \mu(J(r))\, \mathrm{d}r}\, \mathrm{d}s. \tag{1}$$

This will turn out to be a crucial property in this paper.

We are interested in events with an unusual number of jobs in the system. More precisely, we would like to prove an LDP for the number of jobs in the system. A sequence of probability measures $\{\tau_n\}_{n\in\mathbb{N}}$ is said to satisfy an LDP with rate function ρ if there exists a lower semi-continuous function $\rho\colon \mathcal{X} \to [0,\infty]$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \tau_n(F) \leq -\inf_{a\in F} \rho(a)$$

for all closed sets F and

$$\liminf_{n\to\infty} \frac{1}{n} \log \tau_n(G) \geq -\inf_{a\in G} \rho(a)$$

for all open sets G, where each τ_n is defined on the Borel σ-algebra of the topological space $\mathcal{X}$. A sequence of random variables is said to satisfy an LDP with rate function ρ if the sequence of measures induced by the random variables satisfies an LDP with rate function ρ. Importantly, we do not assume that ρ is a good rate function, i.e., we do not assume that ρ has compact level sets.

As mentioned, we would like to prove an LDP for the number of jobs in the system. To analyze this problem, we will scale the arrival rates via $\lambda(x) \mapsto n\lambda(x)$, i.e., we linearly speed up the arrivals. In addition, we will scale the background process via $J \mapsto J_n$. Formally, scaling $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$

means that we start with an infinite-server queue $(J, \lambda, \kappa, \mu)$ and then consider the sequence of infinite-server queues $\{(J_n, n\lambda, \kappa, \mu)\}_{n\in\mathbb{N}}$.

Given the scalings $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$, we denote the corresponding number of jobs in the system by $M_n(t)$. It follows immediately from equation (1) that $M_n(t)$ has a Poisson distribution with random parameter

$$n\phi_t(J_n) = \int_0^t n\lambda(J_n(s))e^{-\kappa(J_n(s))\int_s^t \mu(J_n(r))\,\mathrm{d}r}\,\mathrm{d}s.$$

The normalized random parameter $\phi_t(J_n)$ induces a sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ on $\mathbb{R}$ via $\nu_n(B) = \mathbb{P}(\phi_t(J_n) \in B)$ for Borel sets $B \subset \mathbb{R}$.

We will assume that the sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. Note that $\{\nu_n\}_{n\in\mathbb{N}}$ trivially satisfies an LDP when $\nu_n = \nu_{n+1}$ for all $n \in \mathbb{N}$, so this assumption covers the case in which the background process is not scaled.

Given the scaling, we denote the number of jobs in the system at time t by $M_n(t)$ and consider the normalized random variable $\frac{1}{n}M_n(t)$. Our goal is to prove an LDP for $\frac{1}{n}M_n(t)$ and to describe the corresponding rate function.

Throughout this paper, we will also use the following notation. We denote the closure of a set A by $\mathrm{cl}A$. We write $B(x, \epsilon)$ for the open ball with center $x \in \mathbb{R}^d$ and radius $\epsilon > 0$ and $B[x, \epsilon]$ for its closure. For notational convenience, we will sometimes write $\mathbb{R}_+$ for $[0, \infty)$, $B_+(x, \epsilon)$ for $B(x, \epsilon) \cap \mathbb{R}_+$ and $B_+[x, \epsilon]$ for $B[x, \epsilon] \cap \mathbb{R}_+$. As is customary, we define $\exp(-\infty) = 0$ and $\log(0) = -\infty$.

3 A large deviations principle

In this section we will prove an LDP for the number of jobs in the system under a scaling of the arrival rates and the background process, i.e., we will prove an LDP for $\frac{1}{n}M_n(t)$. It will turn out that so-called attainable parameters determine the rate function corresponding to the LDP.

Definition 3.1. Given a scaling $J \mapsto J_n$, a real number $\gamma \in [0, \infty)$ is called an *attainable parameter* at time $t \geq 0$ if for all $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that $\mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon)) = \nu_n(B(\gamma, \epsilon)) > 0$ for all $n \geq N_\epsilon$. The set of all attainable parameters at time t is denoted by $\mathcal{R}(t)$.

The intuition behind attainable parameters is as follows. The number of jobs in the system has a Poisson distribution with a random parameter that is completely determined by the background process. Basically, the background process samples the Poisson parameter. A real number γ is an attainable parameter if, for all n large enough, the scaled background process samples parameters close to γ with positive probability.

As mentioned before, we will prove an LDP for $\frac{1}{n}M_n(t)$ by scaling $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$ such that the sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. The rate function $I\colon \mathbb{R} \to [0, \infty]$ governing the LDP for $\frac{1}{n}M_n(t)$ is given by

$$I(a) = \inf_{\gamma \in \mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)], \tag{2}$$

where $\ell(\gamma; \cdot)$ is the Fenchel-Legendre transform of the Poisson cumulant generating function with parameter γ. It will turn out (cf. Lemma 3.2) that

$$I(a) = \inf_{\gamma \in \mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)] = \inf_{\gamma \in \{\psi < \infty\}} [\ell(\gamma; a) + \psi(\gamma)]. \tag{3}$$

However, we will take the infimum over $\mathcal{R}(t)$ rather than over $\{\psi < \infty\}$ to stress that attainability of parameters is the crucial property for proving the LDP.

Before we can give the proof, we have to settle some technical details. First, it is not immediately clear whether the function I is indeed a rate function or even whether I is well defined. In particular, it is not clear whether $\mathcal{R}(t)$ is a non-empty set. However, the assumption that the sequence $\{\nu_n\}_{n \in \mathbb{N}}$ satisfies an LDP implies that $\mathcal{R}(t)$ is non-empty, as the following lemma shows.

Lemma 3.2. *Let the scaling $J \mapsto J_n$ be such that $\{\nu_n\}_{n \in \mathbb{N}}$ satisfies an LDP with rate function ψ. Then $\mathcal{R}(t)$ is a non-empty closed subset of $[0, \infty)$ and $\{\psi < \infty\} \subset \mathcal{R}(t)$.*

Proof. Suppose that $\gamma \in \mathbb{R} \setminus \mathcal{R}(t)$. Then there exists $\epsilon > 0$ such that for all $n \in \mathbb{N}$ there exists $k_n \in \mathbb{N}$ such that $k_n \geq n$ and $\nu_{k_n}(B(\gamma, \epsilon)) = 0$. This implies that $B(\gamma, \epsilon) \subset \mathbb{R} \setminus \mathcal{R}(t)$, so $\mathcal{R}(t)$ is closed. Moreover, we must have

$$\liminf_{n \to \infty} \frac{1}{n} \log \nu_n(B(\gamma, \epsilon)) = -\infty = -\inf_{a \in B(\gamma, \epsilon)} \psi(a),$$

so $\psi(a) = \infty$ for all $a \in B(\gamma, \epsilon)$. Then $\mathbb{R} \setminus \mathcal{R}(t) \subset \{\psi = \infty\}$ and $\{\psi < \infty\} \subset \mathcal{R}(t)$. The fact that ψ is a rate function implies that $\{\psi < \infty\}$ is non-empty. The statement of the lemma follows immediately. □

From the previous lemma it follows that I is a well defined function. The fact that I is a rate function is implied by Proposition C.5 and the functions ℓ and ψ being rate functions.

The next lemma is a variation on Varadhan's Lemma. Contrary to Varadhan's Lemma, it does not require that a given function f is continuous. Instead, it requires that a weaker condition is fulfilled. We will use this lemma to obtain the large deviations upper bound, by applying it to functions f of the form described in Proposition C.4.

Lemma 3.3. *Let $\mathcal{X}$ be a topological space and let $\{\xi_n\}_{n \in \mathbb{N}}$ be a sequence of measures defined on its Borel σ-algebra. Suppose that $\{\xi_n\}_{n \in \mathbb{N}}$ satisfies an LDP with rate function ϱ. Let $f \colon \mathcal{X} \to [-\infty, 0]$ be a Borel measurable function such that $f^{-1}([a, b])$ is a closed set for all $a, b \in (-\infty, 0]$ satisfying*

$$\sup_{x \in \mathcal{X}} [f(x) - \varrho(x)] \leq a \leq b \leq 0.$$

Then it holds that

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{X}} e^{nf(x)} \xi_n(\mathrm{d}x) \leq \sup_{x\in\mathcal{X}} [f(x) - \varrho(x)].$$

Proof. This follows immediately from [13, Lem. 2.2] □

With these technical details settled, we can prove the following LDP. Its proof exploits two elementary observations. First, conditional on a value of the random parameter $\phi_t(J_n)$, the number of jobs in the system has the same distribution as the number of jobs in the system in the M/M/∞ setting. Second, the number of jobs in the system in the M/M/∞ setting has the same distribution as a sum of i.i.d. Poisson random variables. Combined with some analytical results, these observations enable us to prove the LDP.

Theorem 3.4. *Consider a modulated infinite-server queue $(J, \lambda, \kappa, \mu)$ as described in Section 2. Scale $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$ such that $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. Then the rescaled number of jobs in the system $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I as defined in equation* (2), *so*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in F\right) \leq -\inf_{a\in F} I(a) \tag{4}$$

for any closed set $F \subset \mathbb{R}$ and

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) \geq -\inf_{a\in G} I(a) \tag{5}$$

for any open set $G \subset \mathbb{R}$.

Proof. For $\lambda \geq 0$, let $P_0(\lambda), P_1(\lambda), P_2(\lambda), \ldots$ denote a sequence of i.i.d. random variables that have a Poisson distribution with parameter λ. Let $F \subset \mathbb{R}$ be a closed set and let $G \subset \mathbb{R}$ be an open set.

To prove the upper bound (4), recall that $M_n(t)$ has a Poisson distribution with random parameter $n\phi_t(J_n)$. Then we may write

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in F\right) &= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} \mathbb{P}\left(\frac{1}{n} P_0(n\gamma) \in F\right) \nu_n(\mathrm{d}\gamma) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} P_i(\gamma) \in F\right) \nu_n(\mathrm{d}\gamma) \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} 2e^{n[-\inf_{a\in F} \ell(\gamma;a)]} \nu_n(\mathrm{d}\gamma) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} e^{n[-\inf_{a\in F} \ell(\gamma;a)]} \nu_n(\mathrm{d}\gamma).
\end{aligned}$$

The inequality above is an immediate result of the proof of Cramér's Theorem in $\mathbb{R}$ as provided in [10].

According to Proposition C.4, the function $\gamma \mapsto -\inf_{a\in F} \ell(\gamma; a)$ satisfies the assumptions of Lemma 3.3. Moreover, $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP both in $\mathbb{R}$ and in $[0,\infty)$ with rate function ψ (cf. [10, Lem. 4.1.5]). Hence, we may apply Lemma 3.3 to obtain

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in F\right) &\leq \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} e^{n[-\inf_{a\in F} \ell(\gamma;a)]}\, \nu_n(\mathrm{d}\gamma) \\ &\leq \sup_{\gamma\in[0,\infty)} \left[-\inf_{a\in F} \ell(\gamma; a) - \psi(\gamma) \right] \\ &= -\inf_{a\in F} \inf_{\gamma\in[0,\infty)} [\ell(\gamma; a) + \psi(\gamma)] \\ &= -\inf_{a\in F} \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)] \\ &= -\inf_{a\in F} I(a).\end{aligned}$$

The fact that we only have to consider the infimum over $\mathcal{R}(t)$ follows from Lemma 3.2. This proves the upper bound.

To prove the lower bound (5), let $\lambda \in \mathcal{R}(t)$ and $\epsilon > 0$. Define $\lambda_\epsilon^- = \max\{0, \lambda - \epsilon\}$ and $\lambda_\epsilon^+ = \lambda + \epsilon$. By definition of the set $\mathcal{R}(t)$ there exists N_ϵ such that $\mathbb{P}(\phi_t(J_n) \in B(\lambda, \epsilon)) > 0$ for all $n \geq N_\epsilon$.

Fix $x \in G$. Because G is open, there exists $\delta > 0$ such that $B(x, \delta) \subset G$. Observe that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) &\geq \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta)\right) \\ &\geq \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta)\,;\, \phi_t(J_n) \in B(\lambda,\epsilon)\right) \\ &= \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta) \,\middle|\, \phi_t(J_n) \in B(\lambda,\epsilon)\right) \mathbb{P}(\phi_t(J_n) \in B(\lambda,\epsilon))\end{aligned}$$

for all $n \geq N_\epsilon$, where the equality follows from the fact that $\mathbb{P}(\phi_t(J_n) \in B(\lambda, \epsilon)) > 0$ for all $n \geq N_\epsilon$. Then we get

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) \geq \\ \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta) \,\middle|\, \phi_t(J_n) \in B(\lambda,\epsilon)\right) \\ + \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \in B(\lambda,\epsilon)).\end{aligned}$$

Recall that $\phi_t(J_n)$ satisfies an LDP with rate function ψ, so

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \in B(\lambda,\epsilon)) \geq -\inf_{a\in B(\lambda,\epsilon)} \psi(a)$$

by assumption. Moreover, it holds that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta) \,\middle|\, \phi_t(J_n) \in B(\lambda,\epsilon) \right) &= \\
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x,\delta) \,\middle|\, \phi_t(J_n) \in B(\lambda,\epsilon) \cap \mathbb{R}_+ \right) &\geq \\
\liminf_{n\to\infty} \inf_{\gamma \in B(\lambda,\epsilon)\cap\mathbb{R}_+} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} P_i(\gamma) \in B(x,\delta) \right) &= \\
\min_{\gamma \in \left\{\lambda_\epsilon^-, \lambda_\epsilon^+\right\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right]&.
\end{aligned}$$

The equality above is established in Proposition C.3. Combining the results, we obtain that

$$\mathbb{P}\left(\frac{1}{n} M_n(t) \in G \right) \geq \min_{\gamma \in \left\{\lambda_\epsilon^-, \lambda_\epsilon^+\right\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right] - \inf_{a \in B(\lambda,\epsilon)} \psi(a).$$

This holds for all $\epsilon > 0$ and small enough $\delta > 0$. Taking limits, we get

$$\lim_{\epsilon \downarrow 0} \min_{\gamma \in \left\{\lambda_\epsilon^-, \lambda_\epsilon^+\right\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right] = - \inf_{a \in B(x,\delta)} \ell(\lambda; a)$$

thanks to Proposition C.4 and

$$\lim_{\epsilon \downarrow 0} \inf_{a \in B(\lambda,\epsilon)} \psi(a) = \psi(\lambda),$$

because ψ is lower semi-continuous. Similarly, we get $\lim_{\delta\downarrow 0} \inf_{a\in B(x,\delta)} \ell(\lambda; a) = \ell(\lambda; x)$. Hence, it follows that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} M_n(t) \in G \right) &\geq \lim_{\delta\downarrow 0} \lim_{\epsilon\downarrow 0} \left[\min_{\gamma \in \left\{\lambda_\epsilon^-, \lambda_\epsilon^+\right\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right] - \inf_{a \in B(\lambda,\epsilon)} \psi(a) \right] \\
&= -[\ell(\lambda; x) + \psi(\lambda)].
\end{aligned}$$

Since $x \in G$ and $\lambda \in \mathcal{R}(t)$ were arbitrary, we obtain

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} M_n(t) \in G \right) &\geq \sup_{a\in G} \sup_{\lambda \in \mathcal{R}(t)} [-[\ell(\lambda; x) + \psi(\lambda)]] \\
&= - \inf_{a \in G} I(a),
\end{aligned}$$

which completes the proof. □

The proof of Theorem 3.4 contains familiar elements. First, the upper bound is proved using a Chernoff bound combined with a variation on Varadhan's Lemma. Second, the lower bound is proved by considering 'the most likely of all unlikely scenarios', which is similar to the method used in [3] and [5]. However, the proofs there relied on properties of irreducible continuous-time Markov chains and the computation of optimal paths, whereas we consider general càdlàg background processes via attainable parameters.

4 Examples: unscaled background processes

Given the scaling $\lambda \mapsto n\lambda$ and $J \mapsto J_n$, Theorem 3.4 provides a full LDP for $\frac{1}{n}M_n(t)$ and describes the corresponding rate function. In the upcoming examples we will consider cases in which the background process is not scaled and we will use Theorem 3.4 to verify or extend known results and to obtain new results.

Throughout this section we will assume that the background process is not scaled, i.e., $J_n = J$ for all $n \in \mathbb{N}$ for some càdlàg stochastic process J. The following lemma is trivial, but plays a central role in this section.

Lemma 4.1. *If $J_n = J$ for all $n \in \mathbb{N}$, then the sequence $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. In this case it holds that $\mathcal{R}(t) = \{\psi < \infty\} = \{\psi = 0\}$.*

Hence, when the background process is not scaled, we have the special property that $\mathcal{R}(t) = \{\psi = 0\}$. This will enable us to compute explicit rate functions in the examples. In these computations, we will extensively use the following properties of the rate function I and properties of step functions in Skorokhod space.

Recall that the rate function I is given by

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)],$$

and that $\mathcal{R}(t) = \{\psi = 0\}$ (see Lemma 4.1). Hence, we get

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} \ell(\gamma; a). \tag{6}$$

In this case, we can give a simpler and more explicit description of I, using the following properties of the function ℓ.

For $\gamma \geq 0$, the function $\ell(\gamma; \cdot)$ is the Fenchel-Legendre transform of the Poisson cumulant generating function with parameter γ and is given by

$$\ell(\gamma; a) = \begin{cases} \infty & a < 0; \\ \gamma & a = 0; \\ \gamma - a + a\log(a/\gamma) & a > 0. \end{cases} \tag{7}$$

For $\gamma = 0$ and $a > 0$, we understand that $\gamma - a + a\log(a/\gamma) = \infty$. An important observation is that the following inequalities hold for $0 \leq \gamma_1 \leq \gamma_2 < \infty$:

$$\ell(\gamma_1; a) \leq \ell(\gamma_2; a) \qquad \forall a \in [0, \gamma_1]; \tag{8}$$

$$\ell(\gamma_1; a) \geq \ell(\gamma_2; a) \qquad \forall a \in [\gamma_2, \infty). \tag{9}$$

See Figure 1 for an illustration.

Because in the present case I is just an infimum of Poisson rate functions, these inequalities imply that I has some special properties. They are described in the following proposition.

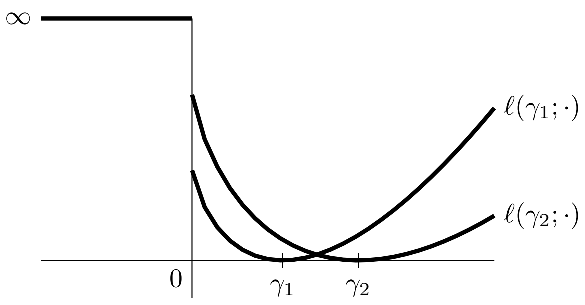

Figure 1: Graphs of the functions $\ell(\gamma_1;\cdot)$ and $\ell(\gamma_2;\cdot)$ for $0 < \gamma_1 < \gamma_2 < \infty$

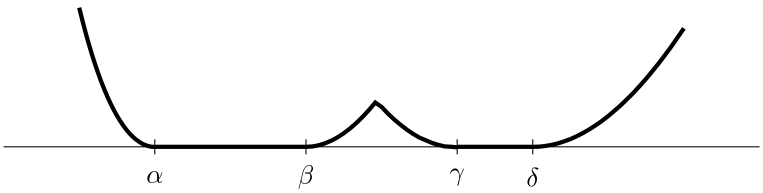

Figure 2: Visualization of the function I in Example 4.3

Proposition 4.2. *In the present case, $I(a) = 0$ if and only if $a \in \mathcal{R}(t)$. If $I(a) > 0$ for some $a \in \mathbb{R}$, then exactly one of the following three scenerios is true:*

1. *$a < c_- = \inf \mathcal{R}(t)$ and $I(b) = \ell(c_-; b)$ for all $b \in (-\infty, c_-]$;*
2. *$a > c_+ = \sup \mathcal{R}(t)$ and $I(b) = \ell(c_+; b)$ for all $b \in [c_+, \infty)$;*
3. *the previous two cases do not hold and $I(b) = \min\{\ell(c_-; b), \ell(c_+; b)\}$ for all $b \in [c_-, c_+]$, where $c_- = \sup(\mathcal{R}(t) \cap (-\infty, a))$ and $c_+ = \sup(\mathcal{R}(t) \cap (a, \infty))$.*

Proof. It follows immediately from equations (6) and (7) that $I(a) = 0$ if and only if $a \in \mathcal{R}(t)$. Hence, $I(a) > 0$ implies that the distance of a to $\mathcal{R}(t)$ is strictly positive, since $\mathcal{R}(t)$ is closed. The three scenarios now follow from the inequalities (8) and (9). □

The previous proposition may seem rather abstract. To get some intuition, the following example describes a typical rate function.

Example 4.3. Suppose that $\mathcal{R}(t) = [\alpha, \beta] \cup [\gamma, \delta]$ for some $0 < \alpha < \beta < \gamma < \delta < \infty$. Then the function I looks like the graph shown in Figure 2: it equals 0 on the intervals $[\alpha, \beta]$ and $[\gamma, \delta]$, whereas it equals the minimum of $\ell(\beta;\cdot)$ and $\ell(\gamma;\cdot)$ on the interval (β, γ) in between. On the interval $(-\infty, \alpha]$ the function I equals $\ell(\alpha;\cdot)$ and on the interval $[\delta, \infty)$ the function I equals $\ell(\delta;\cdot)$.

To compute $\mathcal{R}(t)$, it is often convenient to use the following properties of step functions in $D([0,\infty);\mathcal{E})$. (For the definition of a step function, see Section B.) Recall that the set of all step functions in $D([0,\infty);\mathcal{E})$ is denoted by $\mathcal{S}([0,\infty);\mathcal{E})$.

Lemma 4.4. *If* $\{\phi_t(f) | f \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$*, then*

$$\mathcal{R}(t) = \mathrm{cl}\{\phi_t(f) | f \in \mathcal{S}([0,\infty);\mathcal{E})\} = \{\phi_t(f) | f \in \mathcal{D}([0,\infty);\mathcal{E})\}.$$

Proof. This follows from Lemma 3.2, Corollary B.3 and Lemma B.4. □

Lemma 4.5. *If* $\{\phi_t(f) | f \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$*, then* $\mathcal{R}(t)$ *is a closed interval.*

Proof. It suffices to show that $\mathcal{R}(t)$ is convex. Let $f_{\mathrm{c}}^1, f_{\mathrm{c}}^2 \in \mathcal{S}([0,\infty);\mathcal{E})$. We may assume that $\phi_t\big(f_{\mathrm{c}}^1\big) \leq \phi_t\big(f_{\mathrm{c}}^2\big)$. For $x \in [0,t]$ we define the function g_x via

$$g_x(s) = \mathbb{1}_{\{s<x\}} f_{\mathrm{c}}^1(s) + \mathbb{1}_{\{s \geq x\}} f_{\mathrm{c}}^2(s)$$

for $s \in [0,\infty)$. Clearly, $g_x \in \mathcal{S}([0,\infty);\mathcal{E})$ and

$$\phi_t(g_x) = \phi_x\big(f_{\mathrm{c}}^1\big) + \big(\phi_t\big(f_{\mathrm{c}}^2\big) - \phi_x\big(f_{\mathrm{c}}^2\big)\big).$$

Using the continuity of the integral and applying the Intermediate Value Theorem, it follows that

$$\big[\phi_t\big(f_{\mathrm{c}}^1\big), \phi_t\big(f_{\mathrm{c}}^2\big)\big] \subset \{\phi_t(g_x) | x \in [0,t]\} \subset \mathcal{R}(t).$$

Combined with Lemma 4.4, this implies the statement of the lemma. □

Let $f_{\mathrm{c}} \in \mathcal{S}([0,\infty);\mathcal{E})$ be a step function. Clearly, f_{c} has a unique minimal representation $\{(t_i,\alpha_i)\}_{i=0}^k$, where $k \in \mathbb{N}$, $0 = t_0 < t_1 < \ldots < t_k < \infty$ and $\alpha_0, \ldots, \alpha_k \in \mathcal{E}$ are such that $f_{\mathrm{c}}(t) = \alpha_i$ for $t \in [t_i, t_{i+1})$ and $i = 0, \ldots, k-1$ and $f_{\mathrm{c}}(t) = \alpha_k$ for $t \in [t_k, \infty)$. Given this minimal representation, we define its truncated minimal step size by

$$\Delta_{f_{\mathrm{c}}} = 1 \wedge \min_{i=1,\ldots,k} \{t_i - t_{i-1}\}.$$

Additionally, we define $t_{k+1} = t_k \vee t$. The truncated minimal step size and t_{k+1} will be used for computing attainable parameters.

In the upcoming examples, we would like to compute rate functions via attainable parameters. To compute attainable parameters, we use the following strategy. We fix a certain path f, often a step function. This gives us a parameter value $\phi_t(f)$. Then we would like to show that, with positive probability, the background process stays 'close' to f, which will imply that $\phi_t(f)$ is an attainable parameter.

Staying 'close' to f depends on properties of $\mathcal{E}$ and the background process. In most cases, the background process needs a little bit of room (both in time and in space) to jump near a discontinuity of f. This is where the truncated minimal step size comes in: it is an upper bound on the time we give the background process for jumping near a discontinuity of a step function. The precise meaning of this will become clearer in the examples.

The first example treats the familiar case of a Markov-modulated infinite-server queue, i.e., the case in which the background process is an irreducible Markov chain. This case is partly studied in [3] (Model I) and [5] (Model II). In the example, we recover [3, Th. 2] and [5, Th. 1]. Additionally, we generalize these results to our model and extend them to a full LDP.

Example 4.6. Let J be an irreducible, continuous-time Markov process with finite state space $\mathcal{E} = \{1, \dots, d\}$ and consider the modulated infinite-server queue $(J, \lambda, \kappa, \mu)$. Given the scaling $\lambda \mapsto n\lambda$, Theorem 3.4 (combined with Lemma 4.1) shows that $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I. This rate function may be computed as follows.

Note that $D([0,\infty);\mathcal{E}) = \mathcal{S}([0,\infty);\mathcal{E})$, since $\mathcal{E}$ is finite. Fix any function $g \in D([0,\infty);\mathcal{E})$ with minimal representation $\{(t_i, \alpha_i)\}_{i=0}^{k}$ and take any $\epsilon \in (0,1)$.

Define $\mathcal{W}(g;\epsilon)$ as the set of all $f \in D([0,\infty);\mathcal{E})$ such that

$$\begin{aligned} f(t) &= \alpha_{i-1} \qquad && \forall t \in \left[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\right) \quad \forall i \in \{1, \dots, k\}, \\ f(t) &= \alpha_k && \forall t \in [t_k, t_{k+1}]. \end{aligned}$$

Now note that

$$\sup_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \le \phi_t(g) + \epsilon \max_{j \in \{1,\dots,d\}} \lambda(j)$$

and

$$\inf_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \ge \phi_t(g) - \epsilon \max_{j \in \{1,\dots,d\}} \lambda(j),$$

so we can get both the supremum and the infimum arbitrarily close to $\phi_t(g)$ by taking ϵ small enough.

Observe that $\mathbb{P}(J \in \mathcal{W}(g;\epsilon)) > 0$, thanks to the irreducibility of J. Consequently, $\mathcal{R}(t) = \{\phi_t(g) \,|\, g \in D([0,\infty);\mathcal{E})\}$. Then Lemma 4.5 implies that $\mathcal{R}(t)$ is a closed interval. Using that $\mathcal{E}$ is finite, we immediately get

$$\mathcal{R}(t) = [a_-, a_+],$$

where $0 \le a_- \le a_+ < \infty$ with $a_- = \inf_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$ and $a_+ = \sup_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$. Now applying Proposition 4.2, it follows that the rate function I is given by

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, a_+]; \\ \ell(a_+; a) & a \in [a_+, \infty). \end{cases} \tag{10}$$

The result of the previous example depends neither on the initial distribution nor on the transition rate matrix of the irreducible Markov chain. Moreover, the analysis in the previous example implies the following lemma. It shows that we always obtain a good rate function when the background process has a finite state space.

Lemma 4.7. *Let $J^{(1)}$ be a background process with finite state space $\mathcal{E}$ and let $J^{(2)}$ be an irreducible Markov chain with the same state space. Consider the two modulated infinite-server queues $\big(J^{(1)}, \lambda, \kappa, \mu\big)$ and $\big(J^{(2)}, \lambda, \kappa, \mu\big)$. Scaling $\lambda \mapsto n\lambda$, we obtain in both cases an LDP for the number of jobs in the system with corresponding rate functions $I^{(1)}$ and $I^{(2)}$. Then it holds that $I^{(1)}(a) \geq I^{(2)}(a)$ for all $a \in \mathbb{R}$. In particular, both $I^{(1)}$ and $I^{(2)}$ are good rate functions.*

In the next example we will modulate an infinite-server queue by another Markov-modulated infinite-server queue. This setup differs from the setup considered in [3] and [5]. In particular, the state space of the background process is countably infinite, so that we may obtain a rate function that is not good.

Example 4.8. Consider a Markov-modulated infinite-server queue as described in [14], i.e., a Markov-modulated infinite-server queue under the assumptions of Model I. Assume that neither the arrival rates nor the server work rates are identically equal to 0 and that the system starts empty. Let $J(t)$ be the number of jobs in this Markov-modulated infinite-server queue at time $t \geq 0$. Then J is a càdlàg stochastic process and its state space is $\mathcal{E} = \mathbb{Z}_{\geqslant 0}$.

Consider the modulated infinite-server queue $(J, \lambda, \kappa, \mu)$ and impose the scaling $\lambda \mapsto n\lambda$. Then $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I, according to Theorem 3.4 and Lemma 4.1. This rate function may be computed as follows.

Recall that J stays in state $m \in \mathcal{E}$ during $[t, t+\Delta t]$ with positive probability for arbitrarily large Δt. Moreover, because neither the arrival rates nor the server work rates are identically equal to 0, the process J also has the following property. If $J(t) = m_1$ at time $t \geq 0$, then it jumps to state $m_2 \in \mathcal{E}$ during $[t, t+\Delta t]$ with positive probability for arbitrarily small Δt.

Roughly speaking, these two properties mean that the background process is irreducible, in the sense that it can jump to or stay in any state during any time interval we would like. Of course, this is very similar to the Markov chain being irreducible in the previous example. Consequently, our strategy for determining the attainable parameters will be very similar, although there are some subtleties related to the state space being infinite.

Fix any $g \in \mathcal{S}([0,\infty);\mathcal{E})$ with minimal representation $\{(t_i, \alpha_i)\}_{i=0}^{k}$ and take any $\epsilon \in (0,1)$. Let $\mathcal{W}(g;\epsilon)$ denote the set of all $f \in D([0,\infty);\mathcal{E})$ with

$$\begin{aligned} f(t) &= \alpha_{i-1} && \forall t \in \big[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\big) \quad \forall i \in \{1,\dots,k\}, \\ f(t) &= \alpha_k && \forall t \in [t_k, t_{k+1}], \end{aligned}$$

and

$$\begin{aligned} 0 \leq f(t) &\leq \alpha_0 && \forall t \in \big[0, \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\big), \\ \alpha_{i-1} \wedge \alpha_i \leq f(t) &\leq \alpha_{i-1} \vee \alpha_i && \forall t \in \big[t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\big) \\ & && \forall i \in \{0,\dots,k-1\}, \\ \alpha_{k-1} \wedge \alpha_k \leq f(t) &\leq \alpha_{k-1} \vee \alpha_k && \forall t \in \big[t_k - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_k\big]. \end{aligned}$$

Then we have

$$\sup_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \leq \phi_t(g) + \epsilon \max_{i \in \{0,\dots,k\}} \max_{j \in \{0,\dots,\alpha_i\}} \lambda(j)$$

and

$$\inf_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \geq \phi_t(g) - \epsilon \max_{i \in \{0,\dots,k\}} \max_{j \in \{0,\dots,\alpha_i\}} \lambda(j).$$

The two properties of the background process described above imply that $\mathbb{P}(J \in \mathcal{W}(g;\epsilon)) > 0$. It follows that $\{\phi_t(g) \,|\, g \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$. Write $a_- = \inf_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$ and $a_+ = \sup_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$. Lemma 4.4 and Lemma 4.5 imply that $\mathcal{R}(t) = [a_-, a_+]$ if $a_+ < \infty$ and $\mathcal{R}(t) = [a_-, \infty)$ if $a_+ = \infty$. Hence,

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, a_+]; \\ \ell(a_+; a) & a \in [a_+, \infty) \end{cases} \tag{11}$$

if $a_+ < \infty$ and

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, \infty) \end{cases} \tag{12}$$

if $a_+ = \infty$. Note that I is not a good rate funtion if $a_+ = \infty$.

The previous example only depends on the state space being countable and discrete and on the background process being irreducible in the sense described above. Consequently, the same result holds for irreducible Markov processes with a countable, discrete state space.

In the last example of this section we compare rate functions that are obtained using two different background processes. One background process is a Markov chain, whereas the other background process is a reflected Brownian motion, which has an uncountable state space. It turns out that both background processes lead to the same LDP, even though the background processes are completely different. Apparently, two very different modulating processes may lead to the same rate function for the LDP, even if the arrival rates, service requirements and server work rates are nontrivial.

Example 4.9. Let $\mathcal{E} = [0,1]$ be equipped with the Euclidean metric. Define $\lambda\colon [0,1] \to [0,1]$ by $\lambda(x) = x$, $\kappa\colon [0,1] \to [0,1]$ by $\kappa(x) = 1$ and $\mu\colon [0,1] \to [0,1]$ by $\mu(x) = 1 - x$.

Let J^{MC} be an irreducible, continuous-time Markov chain with state space $\{0,1\}$. Let J^{rBM} be a reflected Brownian motion with reflecting barriers 0 and 1. For simplicity, assume that J^{rBM} starts in $x_0 \in (0,1)$, so

$$J^{\mathrm{rBM}}(t) = x_0 + W(t) + L(t) - U(t)$$

for some standard Brownian motion W, lower-regulator process L and upper-regulator process U (see for instance [8]).

Consider the two modulated infinite-server queues $\big(J^{\mathrm{MC}}, \lambda, \kappa, \mu\big)$ and $\big(J^{\mathrm{rBM}}, \lambda, \kappa, \mu\big)$. Under the scaling $\lambda \mapsto n\lambda$, both $\frac{1}{n} M_n^{\mathrm{rBM}}(t)$ and $\frac{1}{n} M_n^{\mathrm{MC}}(t)$ satisfy an LDP with the same good rate function I, which is given by

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ 0 & a \in [0, t]; \\ \ell(t; a) & a \in [t, \infty). \end{cases} \tag{13}$$

The rate function for the LDP corresponding to $\frac{1}{n} M_n^{\mathrm{MC}}(t)$ is derived in Example 4.6. It is easy to see that the rate function has the form claimed above.

We will show that $\frac{1}{n} M_n^{\mathrm{rBM}}(t)$ satisfies an LDP with the same rate function. Fix $g \in \mathcal{S}([0, \infty); \mathcal{E})$ with minimal representation $\{(t_i, \alpha_i)\}_{i=0}^k$ and take any $\epsilon > 0$. Define $\mathcal{W}(g; \epsilon)$ as the set of all $f \in D([0, \infty); \mathcal{E})$ such that

$$|f(t) - \alpha_i| \leq \epsilon \quad \forall t \in \big[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\big) \quad \forall i \in \{1, \ldots, k\}.$$

Then we get

$$\sup_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \leq \phi_t(g) + \epsilon t + \epsilon$$

and

$$\inf_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \geq \phi_t(g) - \epsilon t - \epsilon.$$

Now observe that

$$\mathbb{P}\big(J^{\mathrm{rBM}} \in \mathcal{W}(g; \epsilon)\big) \geq \mathbb{P}(x_0 + W \in \mathcal{W}(g; \epsilon)) > 0,$$

due to the definition of J^{rBM} and W being a Brownian motion.

It follows that $\{\phi_t(g) \,|\, g \in \mathcal{S}([0, \infty); \mathcal{E})\} \subset \mathcal{R}^{\mathrm{rBM}}(t)$, so $\mathcal{R}^{\mathrm{rBM}}(t) = [0, t]$ and the corresponding rate function is given by the function I above.

In this section we considered examples in which the background process was not scaled. As shown, this implies some special properties, which we can use to explicitly compute rate functions. In the next section, we will scale the background process, too. Although explicit computations are not possible in general, there are still cases for which we may derive rate functions.

5 Examples: scaled background processes

In this section we will give two examples in which the background process is scaled. In the first example, we will consider the Markov-modulated infinite-server queue and derive an explicit rate function under a superlinear time-scaling. In the second example, we will consider a new model in which the background process is a Brownian motion. In this case, the rate function will be given as the solution of a variational problem.

Example 5.1. Let J be an irreducible continuous-time Markov chain with finite state space $\{1,\dots,d\}$ and generator matrix Q. Denote the corresponding stationary distribution by $\pi = (\pi_1,\dots,\pi_d)$.

Consider the modulated infinite-server queue (J,λ,κ,μ). Define $\mu_\infty = \sum_{j=1}^d \pi_j\mu_j$ and

$$\varrho_t = \sum_{j=1}^d \pi_j\lambda_j \int_0^t e^{-\kappa_j\mu_\infty(t-s)}\,\mathrm{d}s = \sum_{j=1}^d \pi_j \frac{\lambda_j}{\kappa_j\mu_\infty}\big(1-e^{-\kappa_j\mu_\infty t}\big).$$

Scale $\lambda \mapsto n\lambda$ and $J \mapsto J_n$, where $J_n(t) = J\big(n^{1+\epsilon}t\big)$. It is easy to see that scaling $J \mapsto J_n$ is equivalent to scaling $Q \mapsto n^{1+\epsilon}Q$.

The sequence of random parameters $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ, where

$$\psi(a) = \begin{cases} 0 & a = \varrho_t; \\ \infty & a \neq \varrho_t. \end{cases}$$

Indeed, this follows from the fact that

$$\liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}(\phi_t(J_n)\in B(\rho_t,\eta)) = 0$$

and

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}(\phi_t(J_n)\notin B(\rho_t,\eta)) = -\infty$$

for all $\eta > 0$. These equalities are an immediate result from the proof of [3, Th. 3].

Given this LDP for $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$, Theorem 3.4 implies that $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I, where

$$I(a) = \ell(\varrho_t;a).$$

Hence, under this superlinear time-scaling of the background Markov chain, the LDP for $\frac{1}{n}M_n(t)$ is governed by a Poisson rate function with parameter ϱ_t.

Example 5.2. Consider a modulated infinite-server queue (J,λ,κ,μ), where the background process J is a standard Brownian motion W on $[0,\infty)$. By $\overline{W}$ we denote its restriction to the interval $[0,t]$. The sample paths of $\overline{W}$ are elements of $C_0[0,t]$, the space of continuous functions $f\colon [0,t]\to\mathbb{R}$ with $f(0)=0$.

Equip $C_0[0,t]$ with the supremum metric. Of course, we may view the function ϕ_t as a map from $C_0[0,t]$ to $[0,\infty)$ and this map is continuous under the supremum metric.

Scale $\lambda \mapsto n\lambda$ and $J \mapsto J_n$, where J_n is given by a linear time-scaling: $J_n(s) = W(s/n)$ for $s \geq 0$. Under this scaling, the arrivals are sped up linearly, whereas the time scale of the Brownian motion is slowed down linearly.

This scaling resembles the scaling featured in [9]. There, the authors considered a modulated infinite-server queue under a linear scaling of both the arrival rate and the time scale of an irreducible Markov chain. The rate function obtained in [9] is given as the solution of a variational problem. We will obtain a similar result in this example.

Since W is a Brownian motion, we have

$$\phi_t(J_n) \overset{\mathrm{d}}{=} \phi_t\left(\frac{1}{\sqrt{n}}W\right) = \phi_t\left(\frac{1}{\sqrt{n}}\overline{W}\right).$$

Schilder's Theorem (cf. [10, Th. 5.2.3]) states that $\frac{1}{\sqrt{n}}\overline{W}$ satisfies an LDP in $C_0[0,t]$ with good rate function

$$\xi(f) = \begin{cases} \frac{1}{2}\int_0^t |\dot{f}(s)|^2 \,\mathrm{d}s & f \in H_1([0,t]); \\ \infty & \text{else.} \end{cases}$$

Here, $H_1([0,t])$ denotes the set of all absolutely continuous functions $f \in C_0[0,t]$ that have square integrable derivative $\dot{f}$.

The contraction principle (cf. [10, Th. 4.2.1]) now implies that $\phi_t(J_n)$ satisfies an LDP with good rate function ψ, where ψ is given by

$$\psi(a) = \inf\{\xi(f) | f \in H_1([0,t]), \phi_t(f) = a\}.$$

It follows from Theorem 3.4 that $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I, where I is given by

$$I(a) = \inf_{\gamma \in \mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)].$$

Now recall that $\{\psi < \infty\} \subset \mathcal{R}(t)$. Also observe that $\{\xi < \infty\} = H_1([0,t])$ and that $\{\psi < \infty\} = \{\phi_t(f) | f \in H_1([0,t])\}$. Then we may rewrite I as

$$\begin{aligned} I(a) &= \inf_{\gamma \in \{\psi < \infty\}} [\ell(\gamma; a) + \psi(\gamma)] \\ &= \inf_{f \in H_1([0,t])} [\ell(\phi_t(f); a) + \psi(\phi_t(f))]. \end{aligned}$$

Hence, I is given as the solution of a variational problem.

6 Discussion and concluding remarks

In this paper, we studied an infinite-server queue in a random environment and proved a full LDP for the transient number of jobs in the system. The proof of this LDP has two essential ingredients, namely the result that the transient number of jobs in the system has a Poisson distribution with a random parameter and the assumption that the random parameter satisfies an LDP. Hence, the large deviations behavior of the random parameter seems to be the crucial factor that determines the large deviations behavior of the number of jobs in the system.

The rate function corresponding to the LDP for the number of jobs is rather abstract. Nevertheless, we showed in the examples how to compute the rate function in certain specific cases. In particular, we recovered earlier obtained results for Markov-modulated infinite-server queues and strengthened these to a full LDP. Additionally, we proved LDPs when the background process has an uncountable state space. In all examples, knowledge about the behavior of the background process could be exploited to describe the rate function.

There are several interesting topics for future research on the modulated infinite-server queue presented here. In this paper, we only looked at large deviations of the number of jobs at a fixed time $t \geq 0$. However, for certain applications it may be desirable to know the deviations over the whole time interval $[0, t]$. Therefore, it would be interesting to consider sample path large deviations. Also moderate deviations could be worth investigating, so as to bridge the gap between the central limit theorems and the large deviations results for modulated infinite-server queues.

Furthermore, it could be very interesting to study other models. In particular, more general arrival processes or service time distributions could be considered. As an example, it seems that the setup of [5] could be generalized to include a background process with countable state space and general service time distributions. Regarding general service time distributions, it should be mentioned that there might be some measurability issues when the background process has an uncountable state space. Finally, it would be interesting to see whether the large deviations results for modulated infinite-server queues carry over to modulated Ornstein-Uhlenbeck processes. To the best of our knowledge, this has not been investigated so far.

Acknowledgement. This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

A Transient number of jobs in the system

In this section, we provide the precise mathematical description of the model and determine the distribution of the number of jobs in the system at time $t \geq 0$, which is denoted by $M(t)$. We mentioned in Section 1 that the steady-state distribution of the number of jobs in the system has already been determined for specific background processes in Model I and Model II. However, in this case we would like to determine the transient distribution given a general background process for the model described below, which generalizes Model I and Model II. Fortunately, the setup of our model is quite convenient and we may obtain the transient distribution without too much effort.

By $D([0,\infty);\mathcal{E})$ we denote the space of càdlàg functions from $[0,\infty)$ to $\mathcal{E}$, where $\mathcal{E}$ is a metric space with metric ρ. We define, in the usual way, a metric

d° on $D([0,\infty);\mathcal{E})$ that generates the Skorokhod J_1 topology. (For more details, see Section B and references there.)

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space on which we have defined an independent, standard Poisson processes $\overline{Y}$ and an independent, càdlàg stochastic process J with state space $\mathcal{E}$. Assume that we have defined a collection of independent standard exponential random variables $\overline{Z}_1, \overline{Z}_2, \ldots$ on this probability space.

To modulate the infinite-server queue, we take continuous functions $\lambda\colon \mathcal{E} \to [0,\infty)$, $\mu\colon \mathcal{E} \to [0,\infty)$ and $\kappa\colon \mathcal{E} \to [0,\infty)$. More precisely, $\lambda(J)$ modulates the arrival rate, $\kappa(J)$ modulates the service requirement distribution and $\mu(J)$ modulates the server work rate.

We define the modulated Poisson process Y via

$$Y(t) = \overline{Y}\biggl(\int_0^t \lambda(J(s))\,\mathrm{d}s\biggr).$$

The process Y will be the arrival process. We denote the jump times of Y by $\tau_1, \tau_2, \ldots$ and the jump times of $\overline{Y}$ by $\overline{\tau}_1, \overline{\tau}_2, \ldots$. For convenience, we set $\tau_0 = \overline{\tau}_0 = 0$. The jump times τ_k and $\overline{\tau}_k$ are related via $\tau_k = \Lambda^-(\overline{\tau}_k)$ and $\overline{\tau}_k = \Lambda(\tau_k)$, where $\Lambda(t) = \int_0^t \lambda(J(s))\,\mathrm{d}s$ and $\Lambda^-(r) = \inf\{t \geq 0 \,|\, \Lambda(t) \geq r\}$.

Define the interarrival times $\sigma_k = \tau_k - \tau_{k-1}$ and $\overline{\sigma}_k = \overline{\tau}_k - \overline{\tau}_{k-1}$ for $k \in \mathbb{N}$. For later use, we note that $\overline{\sigma}_1, \overline{\sigma}_2, \ldots$ is a sequence of i.i.d. random variables with a standard exponential distribution.

At time $t = 0$ there are no jobs in the system. At each jump time of Y exactly one job arrives. Hence, the number of jobs that have entered the system during the time interval $[0,t]$ is given by the (a.s. finite) random variable $\sum_{k=1}^\infty \mathbb{1}_{\{\tau_k \leq t\}}$.

When job k enters the system at time τ_k, it draws an independent service requirement from an exponential distribution with parameter $\kappa(J(\tau_k)) \geq 0$, i.e., the service requirement of job k is given by Z_k, where

$$Z_k = \begin{cases} \overline{Z}_k / \kappa(J(\tau_k)) & \text{if } \kappa(J(\tau_k)) > 0; \\ \infty & \text{if } \kappa(J(\tau_k)) = 0. \end{cases}$$

Job k leaves the system when its service requirement has been processed by the server, whose work rate is modulated by the background process J and is equal to $\mu(J(s))$ for $s \geq 0$.

Hence, job k has both entered and left the system before time $t \geq 0$ if and only if $\tau_k \leq t$ and $Z_k \leq \int_{[\tau_k,t)} \mu(J(r))\,\mathrm{d}r$. We get

$$M(t) = \sum_{k=1}^\infty \biggl(\mathbb{1}_{\{\tau_k \leq t\}} - \mathbb{1}_{\{\tau_k \leq t\}} \mathbb{1}_{\bigl\{Z_k \leq \int_{[\tau_k,t)} \mu(J(r))\,\mathrm{d}r\bigr\}} \biggr).$$

Note that $M(t)$ is a càdlàg stochastic process. Because each Z_k is strictly positive with probability 1, it follows that

$$\begin{aligned} M(t) &\overset{\mathrm{d}}{=} \sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k\le t\}} - \mathbb{1}_{\left\{Z_k<\int_{t\wedge\tau_k}^{t}\mu(J(r))\,\mathrm{d}r\right\}}\right) \\ &= \sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k\le t\}} - \mathbb{1}_{\left\{\overline{Z}_k<\kappa(J(\tau_k))\int_{t\wedge\tau_k}^{t}\mu(J(r))\,\mathrm{d}r\right\}}\right). \end{aligned}$$

If J is deterministic, then it is relatively easy to determine the distribution of $M(t)$. For instance, one may compute the characteristic function of $M(t)$ via the following steps.

Suppose that $J(\omega,t) = f(t)$ for all $\omega \in \Omega$ and $t \ge 0$ for some function $f \in D([0,\infty);\mathcal{E})$. For fixed κ, μ, f and t we define the functions g and h via

$$g(s) = \kappa(f(s))\int_{t\wedge s}^{t}\mu(f(r))\,\mathrm{d}r, \qquad h(s) = 1 + [\exp(\mathrm{i}\theta)-1]\exp(-g(s)).$$

Now we may write the characteristic function of $M(t)$ as

$$\begin{aligned} &\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \mathbb{E}\exp\left(\mathrm{i}\theta\sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k\le t\}} - \mathbb{1}_{\left\{\overline{Z}_k<\kappa(f(\tau_k))\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right\}}\right)\right) = \\ &= \mathbb{E}\mathbb{1}_{\{\tau_1>t\}} + \sum_{n=1}^{\infty}\mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\exp\left(\mathrm{i}\theta\left(n-\sum_{k=1}^{n}\mathbb{1}_{\left\{\overline{Z}_k<g(\tau_k)\right\}}\right)\right). \end{aligned}$$

Clearly, $\mathbb{E}\mathbb{1}_{\{\tau_1>t\}} = e^{-\int_0^t\lambda(f(s))\,\mathrm{d}s} = e^{-\Lambda(t)}$. We are left with computing the infinite sum above. Fix $n\in\mathbb{N}$ and note that

$$\begin{aligned} &\mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\exp\left(\mathrm{i}\theta\left(n-\sum_{k=1}^{n}\mathbb{1}_{\left\{\overline{Z}_k<g(\tau_k)\right\}}\right)\right) = \\ &= \mathbb{E}\left(\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\mathbb{E}\left[\exp\left(\mathrm{i}\theta\left(n-\sum_{k=1}^{n}\mathbb{1}_{\left\{\overline{Z}_k<g(\tau_k)\right\}}\right)\right)\middle|\,\tau_1,\tau_2,\dots\right]\right) \\ &= \mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\prod_{k=1}^{n}\left(1+[\exp(\mathrm{i}\theta)-1]e^{-g(\tau_k)}\right), \end{aligned}$$

because Y and $\overline{Z}_1,\overline{Z}_2,\dots$ are independent.

Next, observe that

$$\mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\prod_{k=1}^{n}\left(1+[\exp(\mathrm{i}\theta)-1]e^{-g(\tau_k)}\right) = \mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}}\prod_{k=1}^{n}h(\tau_k)$$

and

$$\mathbb{E}\mathbb{1}_{\{\tau_n \le t; \tau_{n+1} > t\}} \prod_{k=1}^{n} h(\tau_k) = \mathbb{E}\left(\mathbb{1}_{\{\tau_n \le t\}}\left(\prod_{k=1}^{n} h(\tau_k)\right)\mathbb{E}\big[\mathbb{1}_{\{\sigma_{n+1} > t-\tau_n\}} \,\big|\, \tau_1, \dots, \tau_n\big]\right)$$
$$= \mathbb{E}\left(\mathbb{1}_{\{\tau_n \le t\}}\left(\prod_{k=1}^{n} h(\tau_k)\right)e^{-(\Lambda(t)-\Lambda(\tau_n))}\right).$$

For convenience we write $x_k^+ = x_1 + \cdots + x_k$. We have

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{1}_{\{\tau_n \le t\}}\left(\prod_{k=1}^{n} h(\tau_k)\right)e^{\Lambda(\tau_n)}\right) = \\
&= \mathbb{E}\mathbb{1}_{\{\overline{\tau}_n \le \Lambda(t)\}}\left(\prod_{k=1}^{n} h\big(\Lambda^-(\overline{\tau}_k)\big)\right)e^{\overline{\tau}_n} \\
&= \int_{x_1=0}^{\Lambda(t)} \int_{x_2=0}^{\Lambda(t)-x_1^+} \cdots \int_{x_n=0}^{\Lambda(t)-x_{n-1}^+} \prod_{k=1}^{n} h\big(\Lambda^-\big(x_k^+\big)\big)\,\mathrm{d}x_n \dots \mathrm{d}x_1 \\
&= \int_{y_1=0}^{\Lambda(t)} \int_{y_2=y_1}^{\Lambda(t)} \cdots \int_{y_n=y_{n-1}}^{\Lambda(t)} \prod_{k=1}^{n} h\big(\Lambda^-(y_k)\big)\,\mathrm{d}y_n \dots \mathrm{d}y_1 \\
&= \int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} [h(z_k)\lambda(f(z_k))]\,\mathrm{d}z_n \dots \mathrm{d}z_1.
\end{aligned}$$

Now note that for an integrable function g we have

$$\left[\int_0^t g(s)\,\mathrm{d}s\right]^n = n! \int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} g(z_k)\,\mathrm{d}z_n \dots \mathrm{d}z_1.$$

As a result, it holds that

$$\begin{aligned}
&\int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} [h(z_k)\lambda(f(z_k))]\,\mathrm{d}z_n \dots \mathrm{d}z_1 = \\
&= \frac{1}{n!}\left[\int_0^t h(s)\lambda(f(s))\,\mathrm{d}s\right]^n \\
&= \sum_{k=0}^{n} \frac{1}{k!}\Lambda(t)^k \frac{1}{(n-k)!}\left([\exp(\mathrm{i}\theta) - 1]\int_0^t \lambda(f(s))e^{-g(s)}\,\mathrm{d}s\right)^{n-k}.
\end{aligned}$$

Now we may write

$$\begin{aligned}
&\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \\
&= \mathbb{E}\mathbb{1}_{\{\tau_1>t\}} + \sum_{n=1}^{\infty} \mathbb{E}\mathbb{1}_{\{\tau_n\le t;\tau_{n+1}>t\}} \exp\Bigg(\mathrm{i}\theta\Bigg(n - \sum_{k=1}^{n} \mathbb{1}_{\{\overline{Z}_k<g(\tau_k)\}}\Bigg)\Bigg) \\
&= e^{-\Lambda(t)} + \sum_{n=1}^{\infty} e^{-\Lambda(t)} \sum_{k=0}^{n} \frac{1}{k!}\Lambda(t)^k \frac{1}{(n-k)!}\Bigg(\big[e^{\mathrm{i}\theta}-1\big]\int_0^t \lambda(f(s))e^{-g(s)}\,\mathrm{d}s\Bigg)^{n-k} \\
&= e^{-\Lambda(t)} \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{1}{k!}\Lambda(t)^k \frac{1}{(n-k)!}\Bigg(\big[e^{\mathrm{i}\theta}-1\big]\int_0^t \lambda(f(s))e^{-g(s)}\,\mathrm{d}s\Bigg)^{n-k} \\
&= e^{-\Lambda(t)} \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{1}{k!}\Lambda(t)^k \frac{1}{n!}\Bigg(\big[e^{\mathrm{i}\theta}-1\big]\int_0^t \lambda(f(s))e^{-g(s)}\,\mathrm{d}s\Bigg)^{n} \\
&= \exp\Bigg([\exp(\mathrm{i}\theta)-1]\int_0^t \lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s\Bigg).
\end{aligned}$$

Hence, in this case $M(t)$ has a Poisson distribution with parameter $\phi_t(f)$, where $\phi_t(f) = \int_0^t \lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s$.

Now suppose that J is not deterministic. Then J is a random element of $D([0,\infty);\mathcal{E})$. In this case, we may use the independence of J and standard arguments to obtain that

$$\begin{aligned}
\mathbb{E}\exp(\mathrm{i}\theta M(t)) &= \mathbb{E}\mathbb{E}\big[\exp(\mathrm{i}\theta M(t))\big|\mathcal{F}_\infty^J\big] \\
&= \mathbb{E}\exp\Bigg([\exp(\mathrm{i}\theta)-1]\int_0^t \lambda(J(s))e^{-\kappa(J(s))\int_s^t \mu(J(r))\,\mathrm{d}r}\,\mathrm{d}s\Bigg).
\end{aligned}$$

We summarize our findings in the following lemma.

Lemma A.1. *Under the stated conditions, $M(t)$ has a Poisson distribution with random parameter*

$$\phi_t(J) = \int_0^t \lambda(J(s))e^{-\kappa(J(s))\int_s^t \mu(J(r))\,\mathrm{d}r}\,\mathrm{d}s.$$

Consequently, if we scale $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$, then the number of jobs in the system $M_n(t)$ has a Poisson distribution with random parameter $n\phi_t(J_n)$. This observation is crucial for the proof of the LDP for $\frac{1}{n}M_n(t)$.

B Continuity in Skorokhod space

Let $\mathcal{E}$ be a metric space with metric ρ. Let $D([0,\infty);\mathcal{E})$ denote the space of càdlàg functions $f\colon [0,\infty) \to \mathcal{E}$, i.e., $\lim_{s\downarrow t} f(s) = f(t)$ and $\lim_{s\uparrow t} f(s)$ exists in $\mathcal{E}$ for every $t \ge 0$, where $\lim_{s\uparrow 0} f(s) := f(0)$ by convention.

Define a metric d° on $D([0,\infty);\mathcal{E})$ via

$$d^\circ(f,g) = \inf_{\lambda\in\Lambda}\left[\gamma(\lambda) \vee \int_0^\infty e^{-u} d(f,g,\lambda,u)\,\mathrm{d}u\right].$$

Here, Λ denotes the space of increasing homeomorphisms of $[0,\infty)$,

$$\gamma(\lambda) = \sup_{t>s\geq 0} |\log(\lambda(t)-\lambda(s)) - \log(t-s)|$$

and

$$d(f,g,\lambda,u) = \sup_{t\in[0,\infty)} [1 \,\wedge\, \rho(f(t\wedge u), g(\lambda(t)\wedge u))].$$

The metric d° induces the Skorokhod J_1 topology. For more details, see [11] or [16].

Definition B.1. A function $f_\mathrm{c} \in D([0,\infty);\mathcal{E})$ is called a *piecewise constant function* or a *step function* if there exist $n \in \mathbb{N}$, finitely many time points $0 = t_0 < t_1 < \ldots < t_n < \infty$ and $\alpha_0,\ldots,\alpha_n \in \mathcal{E}$ such that $f_\mathrm{c}(t) = \alpha_i$ for $t \in [t_i, t_{i+1})$ and $i = 0,\ldots,n-1$ and $f_\mathrm{c}(t) = \alpha_n$ for $t \in [t_n,\infty)$.

The set of step functions in $D([0,\infty);\mathcal{E})$ is denoted by $\mathcal{S}([0,\infty);\mathcal{E})$.

Proposition B.2. *Let $f \in D([0,\infty);\mathcal{E})$. For all $T > 0$ and $\epsilon > 0$ there exists a step function $f_\mathrm{c} \in \mathcal{S}([0,\infty);\mathcal{E})$ such that*

$$\sup_{t\in[0,T]} \rho(f(t), f_\mathrm{c}(t)) < \epsilon.$$

Proof. This is derived in the same way as [16, Th. 12.2.2]. □

Corollary B.3. *The set $\mathcal{S}([0,\infty);\mathcal{E})$ is dense in $D([0,\infty);\mathcal{E})$.*

Consequently, every continuous function on $D([0,\infty);\mathcal{E})$ is completely determined by its behavior on the set of step functions.

Let $\lambda\colon \mathcal{E} \to [0,\infty)$, $\kappa\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous. For $t \geq 0$, we would like to show that the function $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ defined via

$$\phi_t(f) = \int_0^t \lambda(f(s)) e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s \tag{14}$$

is a continuous function.

First, we observe that the map $c_\lambda\colon D([0,\infty);\mathcal{E}) \to D([0,\infty);\mathbb{R})$ defined via $c_\lambda(f)(t) = \lambda(f(t))$ is continuous, because λ is continuous. Similarly, the functions c_κ and c_μ are continuous.

Next, let $f, g \in D([0,\infty);\mathbb{R})$. Then pointwise multiplication of f and g, defined via $(fg)(t) = f(t)g(t)$. This is a measurable map which is continuous at (f,g) if f or g is continuous (cf. [15, Th. 4.2]).

Finally, let $f \in D([0,\infty);\mathbb{R})$. Then the map $\psi\colon D([0,\infty);\mathbb{R}) \to D([0,\infty);\mathbb{R})$ defined via $\psi(t) = \int_0^t f(s)\,\mathrm{d}s$ is continuous. This follows almost immediately from the definition of ψ and the characterization in [11, Pr. 3.5.3].

Now note that the sequence of functions $\{\lambda(f_n)\}_{n\in\mathbb{N}}$ is bounded in the sup norm over $[0,t]$ if $f_n \to f$ in $D([0,\infty);\mathcal{E})$. Hence, it suffices to show that

$$\int_0^t e^{-\kappa(f_n(s))\int_s^t \mu(f_n(r))\,\mathrm{d}r}\,\mathrm{d}s \to \int_0^t e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s$$

as $f_n \to f$ in $D([0,\infty);\mathcal{E})$. But this follows from repeated applications of the first three observations.

Hence, the map ϕ_t must be continuous. Note that continuity of λ, κ and μ is crucial to obtain this result. We summarize these findings in the following lemma.

Lemma B.4. *Let $\lambda\colon \mathcal{E} \to [0,\infty)$, $\kappa\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous. Then the function $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ as defined in equation* (14) *is continuous.*

C Properties of Poisson random variables

For $\gamma \geq 0$, let $P_0(\gamma), P_1(\gamma), P_2(\gamma), \ldots$ denote a sequence of i.i.d. random variables that have a Poisson distribution with parameter γ. In this section, we will fix an arbitrary $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ and define $\lambda_\epsilon^- = \max\{0, \lambda - \epsilon\}$ and $\lambda_\epsilon^+ = \lambda + \epsilon$. Recall that $B_+(\lambda,\epsilon) = B(\lambda,\epsilon) \cap \mathbb{R}_+$.

We would like to prove a large deviations lower bound for

$$\liminf_{n\to\infty} \inf_{\gamma \in B_+(\lambda,\epsilon)} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right).$$

Of course, the difficulty here is the presence of the infimum over a range of parameters. We will show in Proposition C.1 that this infimum may be taken over certain restricted subsets of $B_+(\lambda,\epsilon)$. For each of these subsets we will provide a large deviations lower bound, from which we will derive a lower bound when the infimum is taken over $B_+(\lambda,\epsilon)$. This is the content of Proposition C.3.

Proposition C.1. *For all $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ it holds that*

$$\inf_{\gamma \in B_+(\lambda,\epsilon)} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right) = \inf_{\gamma \in (B(\lambda,\epsilon)\cap B[x,\delta]) \cup \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right).$$

Proof. Let $0 \leq \gamma_- \leq \gamma_+ < \infty$. For $y \in \mathbb{R}$ it holds that

$$\mathbb{P}(P_0(\gamma_+) = y) \geq \mathbb{P}(P_0(\gamma_-) = y) \qquad \text{if} \qquad y \geq \gamma_+ \geq \gamma_- \tag{15}$$

and

$$\mathbb{P}(P_0(\gamma_+) = y) \leq \mathbb{P}(P_0(\gamma_-) = y) \qquad \text{if} \qquad \gamma_+ \geq \gamma_- \geq y. \tag{16}$$

Because we are working with i.i.d. Poisson random variables, we may write

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in B(x,\delta)\right) = \mathbb{P}(P_0(n\gamma) \in (n(x-\delta), n(x+\delta))). \tag{17}$$

Now the statement of the proposition is an easy consequence of the equations (15), (16) and (17) combined. □

Proposition C.2. *Let $x \in \mathbb{R}$ and $\delta > 0$. If $B_+(x,\delta) \neq \emptyset$, then*

$$\lim_{n\to\infty} \inf_{\gamma \in B_+[x,\delta]} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in B(x,\delta)\right) = 0.$$

Proof. For a Borel set $A \subset \mathbb{R}$, define $p_n(A|\gamma) = \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in A\big)$. Now suppose that $B_+(x,\delta) \neq \emptyset$. Then the diameter of $B_+(x,\delta)$ is strictly positive and bounded above by $r = \min\{2\delta, x+\delta\}$.

Let $N_r \in \mathbb{N}$ be such that $\frac{1}{N_r} < \frac{r}{2}$. Then for all $n \geq N_r$ and $\gamma \in B_+[x,\delta]$ we define $\gamma_n^- = \frac{1}{n}\lfloor n\gamma \rfloor$, $\gamma_n^+ = \frac{1}{n}\lceil n\gamma \rceil$ and

$$\gamma_n^* = \min\big\{\{\gamma_n^-, \gamma_n^+\} \cap B(x,\delta)\big\}.$$

Then $\max\{|\gamma - \gamma_n^-|, |\gamma - \gamma_n^+|\} \leq \frac{1}{n} < \frac{r}{2}$ and $p_n(B(x,\delta)|\gamma) \geq p_n(\{\gamma_n^*\}|\gamma)$ for each $n \in \mathbb{N}$ and each $\gamma \in B_+[x,\delta]$. Using that $n! \leq n^{n+1/2}e^{-n+1}$, we get

$$\begin{aligned} p_n(\{\gamma_n^*\}|\gamma) &\geq \left(\frac{n\gamma}{n\gamma+1}\right)^{n\gamma_n^*} e^{n(\gamma_n^*-\gamma)}e^{-1}(n\gamma_n^*)^{-1/2} \\ &\geq \left(1 - \frac{1}{n(x+\delta)+1}\right)^{n(x+\delta)} e^{-2}(n(x+\delta))^{-1/2} \end{aligned}$$

for each $n \in \mathbb{N}$ and each $\gamma \in B_+[x,\delta]$. This implies the statement. □

Combined with Cramér's Theorem in $\mathbb{R}$, the two previous propositions enable us to prove the following large deviations bound. Note that we prove an equality rather than an inequality and that the limit exists.

Proposition C.3. *For all $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ it holds that*

$$\lim_{n\to\infty} \inf_{\gamma \in B_+(\lambda,\epsilon)} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in B(x,\delta)\right) = \min_{\gamma \in \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a)\right]. \tag{18}$$

Proof. Define $p_n(A \,|\, \gamma) = \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in A\big)$ for Borel sets $A \subset \mathbb{R}$ and $C = (B(\lambda, \epsilon) \cap B[x, \delta]) \cup \{\lambda_\epsilon^-, \lambda_\epsilon^+\}$. Thanks to Proposition C.1 we may write

$$\lim_{n\to\infty} \inf_{\gamma \in B_+(\lambda,\epsilon)} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) = \lim_{n\to\infty} \inf_{\gamma\in C} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma).$$

It follows from Proposition C.2 that we may restrict the infimum to the set $\{\lambda_\epsilon^-, \lambda_\epsilon^+\}$, so

$$\begin{aligned}
\lim_{n\to\infty} \inf_{\gamma\in C} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) &= \lim_{n\to\infty} \min_{\gamma\in\{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) \\
&= \min_{\gamma\in\{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \lim_{n\to\infty} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) \\
&= \min_{\gamma\in\{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \left[-\inf_{a\in B(x,\delta)} \ell(\gamma; a) \right].
\end{aligned}$$

The last equality is an application of Cramér's Theorem for i.i.d. Poisson random variables; the limit exists because $B(x, \delta)$ is a continuity set for the Fenchel-Legendre transform corresponding to a Poisson distribution. □

As shown in the inequalities (8) and (9), the Fenchel-Legendre transforms corresponding to Poisson distributions are nicely ordered in some sense. This property leads to the following propositions. Their proofs are elementary but tedious and are therefore omitted.

Proposition C.4. *Let $F \subset \mathbb{R}$ be closed and define $f\colon [0,\infty) \to [-\infty, 0]$ via*

$$f(\gamma) = -\inf_{a\in F} \ell(\gamma; a).$$

If $F \subset (-\infty, 0)$, then $f \equiv -\infty$. If $F \cap [0,\infty) \neq \emptyset$, then f is real-valued and continuous on $(0,\infty)$. Additionally, $\lim_{\gamma\downarrow 0} f(\gamma) = f(0)$, where $f(0) = 0$ if $0 \in F$ and $f(0) = \infty$ if $0 \notin F$. In any case, $f^{-1}([a, b])$ is closed for all $a, b \in (-\infty, 0]$ with $a \leq b$.

Proposition C.5. *Let $\mathcal{R} \subset [0,\infty)$ be a non-empty, closed set. Let $\psi\colon \mathbb{R} \to [0,\infty]$ be a lower semi-continuous function. Then the function $I\colon \mathbb{R} \to [0,\infty]$ defined via*

$$I(a) = \inf_{\gamma\in\mathcal{R}} [\ell(\gamma; a) + \psi(\gamma)]$$

is a lower semi-continuous function.

References

[1] J. D. Biggins. Large deviations for mixtures. *Electronic Communications in Probability*, 9:60–71, 2004.

[2] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir. Markov-modulated infinite-server queues with general service times. *Queueing Systems*, 76(4):403–424, 2014.

[3] Joke Blom, Koen De Turck, Offer Kella, and Michel Mandjes. Tail asymptotics of a Markov-modulated infinite-server queue. *Queueing Systems*, 78(4):337–357, 2014.

[4] Joke Blom, Koen De Turck, and Michel Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. Submitted, 2014.

[5] Joke Blom and Michel Mandjes. A large-deviations analysis of Markov-modulated infinite-server queues. *Operations Research Letters*, 41(3):220–225, 2013.

[6] Pauline Coolen-Schrijner and Erik A. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, 16(3):351–366, 2002.

[7] B. D'Auria. M/M/∞ queues in semi-Markovian random environment. *Queueing Systems*, 58(3):221–237, 2008.

[8] B. D'Auria, J. Ivanovs, O. Kella, and M. Mandjes. Two-sided reflection of Markov-modulated Brownian motion. *Stochastic Models*, 28(2):316–332, 2012.

[9] K. E. E. S. De Turck and M. R. H. Mandjes. Large deviations of an infinite-server system with a linearly scaled background process. *Performance Evaluation*, 75-76:36–49, 2014.

[10] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, second edition, 1998.

[11] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[12] Brian H. Fralix and Ivo J. B. F. Adan. An infinite-server queue influenced by a semi-Markovian environment. *Queueing Systems*, 61(1):65–84, 2009.

[13] H. M. Jansen, M. R. H. Mandjes, K. De Turck, and S. Wittevrongel. On the upper bound in Varadhan's Lemma. arXiv:1411.3568v1, 2014.

[14] C. A. O'Cinneide and P. Purdue. The M/M/∞ queue in a random environment. *Journal of Applied Probability*, 23(1):175–184, 1986.

[15] Ward Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5(1):67–85, 1980.

[16] Ward Whitt. *Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, 2002.